Heredity (2019) 123:58–66
https://doi.org/10.1038/s41437-019-0226-y

the geneticssociety

REVIEW ARTICLE

Genome sequencing—the dawn of a game-changing era

Veronica van Heyningen [1,2]

Received: 4 April 2019 / Accepted: 16 April 2019

Abstract
The development of genome sequencing technologies has revolutionized the biological sciences in ways which could not have been imagined at the time. This article sets out to document the dawning of the age of genomics and to consider the impact of this revolution on biological investigation, our understanding of life, and the relationship between science and society.

Introduction

The congruence of several critical technologies was needed to launch the genome era, allowing mapping, sequencing, assembly and analysis of whole genomes. The discovery of restriction enzymes, development of cloning vectors for making representative genomic libraries, hybridization techniques like DNA (Southern) blotting and PCR (polymerase chain reaction) for easy DNA amplification were all major landmarks. Observation of frequent restriction site polymorphisms for mapping markers gave a new lease of life to classical family linkage studies by hugely expanding the repertoire of individual variation that could be easily assessed. While individual variation is essential for classical genetic analysis, it is not required for all mapping exercises, however, the development of chromosomal in situ hybridization for regional chromosomal assignments in mitotic somatic cells was used for dense regional mapping of DNA clones. The advent of DNA sequencing technology was a huge milestone. It was, of necessity, accompanied by improved data storage and retrieval, and novel analytical approaches to empower sequence comparison. Phenotype, gene, and anonymous marker mapping in different organisms were required to deliver biological context to genome mapping.

Quite apart from its historic status as the first multicellular eukaryotic model organism to be sequenced and assembled in full by 1998, the *Caenorhabditis elegans* genome project should be celebrated as the original paradigm for all the other multicellular genomes. The collaborative ethos developed for its efficient and speedy delivery is the model for everything that followed. Its triumph gave John Sulston the insight and confidence to fight successfully to keep the multi-centre Human Genome Project (HGP) fully in the public domain.

The genomials

My working life as a human geneticist, from the start of my DPhil in 1970, has been lit and fuelled by fantastic advances in technology-driven knowledge. Life as a Genomial (cf. Millennial) remains fast paced and exciting, so much so, that in "retirement" it is proving difficult to relinquish my involvement. It is astonishing to reflect that the earliest genome sequence, for a 5.4 kb bacteriophage, emerged only just over 40 years ago when I was already a postdoc. In fact in 1974, when I joined the MRC Mammalian Genome Unit in Edinburgh, the very word "genome" was not readily understood by the general public (although the term was actually coined in 1920 by Hamburg botanist Hans Winkler). Over the years, several disciplines have been radically redirected by the advent of the genome era, and this will continue as technological advances provide ever more ambitious opportunities. As the centenary year of the Genetics Society arrives, the international biology community is planning the sequencing of at least one exemplar of all known species on the globe, as part of the Earth Biogenome Project [https://www.earthbiogenome.org/].

✉ Veronica van Heyningen
veronica.vanheyningen@igmm.ed.ac.uk

[1] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Crewe Road, Edinburgh EH4 2XU, UK

[2] Institute of Ophthalmology, University College London, 11-43 Bath Street, London EC1V 9EL, UK

The delivery of sequence is now almost trivial; the challenge lies in interpretation, linking to phenotype, curation and imaginative, ambitious interrogation of the data, and, as we shall see, international communication and collaboration are critical for the fast evolution of all this technology.

Assembling the toolbox and defining aims

What is a genome, what are its functions?

Many coffee-time discussions at MRC MGU in the mid 1970s were about the size and possible organization of genomes and about the likely number of genes in different organisms. It was assumed that gene number would show close correlation with biological complexity. Quantitative nucleic acid hybridization studies predicted many repetitious DNA regions in most genomes. Transcript analysis revealed that only a small proportion of a genome is represented in transcribed RNA. The existence of introns between the exonic sequences of genes was a surprise discovery emerging in 1977 (Berget et al. 1977; Chow et al. 1977). Messenger RNA is processed by splicing out the introns. Many genes give rise to multiple alternative splice isoforms, some of which have critically distinct functions. A significant proportion of most genomes was deduced to be non-coding, and initially labelled "junk" DNA, though maintenance of "useless" DNA is not readily reconciled with concepts of evolution through selection. With the appearance of some of the first complex genomes, particularly the realization that the gene numbers in *C. elegans*, fruit flies and humans are not hugely different, considerable re-thinking of genome function was required. The role of large, complex, non-coding regulatory regions has become accepted. Regulatory functions are critical for spatio-temporal and quantitative control of gene expression in cell proliferation, development and differentiation. Similarly, the spatial organization of the genome within the cell nucleus has been a recognized problem ever since genome sizes were defined. The human haploid genome size is 3.2 Gb, amounting to a length of 2 m in each diploid cell nucleus. Thus, the DNA needs to be much compacted, which requires the help of proteins, producing complex but dynamic chromatin conformations that change as gene expression is turned on or off.

Gene mapping at the protein level

My DPhil in Walter Bodmer's Oxford lab focused on early human gene mapping using somatic cell hybrids (van Heyningen et al. 1973). The use of somatic cells, with Pontecorvo as an early advocate of the approach (Pontecorvo 1971), was particularly useful for human genetics where controlled breeding experiments are not possible and family linkage studies are impeded by small family size. These early studies pre-dated DNA level analysis, relying instead on protein identification methods by function or with specific antibodies. Though novel and productive at the time, this approach had limited scope, as it required that the gene to be mapped is expressed in the rodent-human hybrid cell line used. Fortuitously, such cells, the products of fusion between an established rodent cell line with normal diploid human cells, randomly lose human chromosomes, creating a segregating system suitable for chromosomal gene assignment (Pontecorvo 1971). Cytogeneticists learnt to distinguish and identify the human chromosomes from the more constant rodent contribution in these somatic cell hybrids (Bobrow and Cross 1974). A little later, with the availability of monoclonal antibodies, frequently with species-specific marker recognition, the genes for expressed cell surface markers could also be mapped in hybrid cells (Kao et al. 1977).

Classical family linkage studies were also rejuvenated by the use of the new mapping technique involving expressed genes, such as polymorphic enzyme markers distinguished electrophoretically, and polymorphic cell surface markers such as blood group antigens that could be followed using antibodies. Linkages were identified using logarithm for odds (LOD) scores (Morton 1996). Family studies were also essential for variant and disease phenotype linkage to progressively tighter chromosomal localizations.

The advent of DNA-level analysis

In 1974 when I arrived at MRC MGU, Ed Southern was still perfecting his newly invented DNA blotting technique, to be published the following year (Southern 1975), in the *Journal of Molecular Biology*, a highly favoured destination for key papers at that time. The new technology, "Southern blotting", was used initially to characterize repetitive mouse satellite DNA, the periodicities of which were revealed by cutting with specific restriction enzymes. Satellite DNAs from rodents were identified in the course of density gradient ultra-centrifugation, then used routinely to purify DNA. They were multi-copy repeats with distinct nucleotide composition, which ran as a "satellite" peak, differing in density from the main DNA band. In the cell, these fractions were found near the centromeric regions of chromosomes. Aliquots of recently discovered restriction enzymes (Bigger et al. 1973) came at first as gifts from Noreen Murray's nearby lab, but soon a young biochemist was employed at MRC MGU for the sole task of preparing restriction enzymes for the Unit. In 1979, the human alpha-globin cluster was mapped to chromosome 11, still using somatic cell hybrids, restriction enzyme fragmentation and Southern blotting (Jeffreys et al. 1979). Adoption of the

new techniques permitted faster more generic gene mapping. Any piece of DNA, not just transcribed segments, could be mapped in this way. Gene expression was not required. With the availability of restriction enzymes and cloning vectors ready to receive appropriate DNA fragments, whole organismal genomes could, in principle, be represented in libraries, and individual clones could be systematically assigned to chromosomes and ordered using different technologies.

In parallel with characterizing restriction enzymes, Noreen Murray was also pioneering the development of the earliest cloning vectors, based on lambda phage (Borck et al. 1976). In the early 1980s, a whole host of new vectors with distinct characteristics, and different capacities for inserted DNA, emerged (including cosmids, BACs and YACs—bacterial and yeast artificial chromosomes, respectively) (Anand et al. 1989; Holland et al. 1993; Wang et al. 1994). These and other genomic clones were beginning to be used to identify restriction fragment length polymorphisms (RFLPs) to define haplotypes and population variation with the suggestion that these variants could be used for mapping (Solomon and Bodmer 1979), an idea subsequently elaborated on by others (Botstein et al. 1980). Around this time, cDNA (DNA complementary to mRNAs) libraries were also produced, from many different tissues to help define transcribed functional genes, preferentially. Some considered cDNA mapping to be the logical way to map the genome, since only 1% of the genome falls into this category. I recall an early human genome mapping meeting at the London Zoo, where Sydney Brenner made this suggestion. Short expressed sequence tags (ESTs) were soon used as mapping tools. However, single-nucleotide polymorphisms (SNPs) covered the genome more fully and evenly, and emerged as the most readily used mapping tools (Wang et al. 1998). cDNA clones were still essential to define transcribed genes. Localised mapping studies often used mini- and microsatellite repeat polymorphisms (Jeffreys et al. 1993).

Hunting for "disease genes"

The second stage of my own gene mapping efforts began soon after I became a tenured scientist at the MRC Human Genetics Unit (HGU), searching for genes implicated in congenital disease, in collaboration with Nick Hastie. This was no longer mapping for mapping's sake. We now had an overriding biomedical aim. On 9 June 1983, general election day, after extensive discussion, we chose to work on a "contiguous gene syndrome" where two diseases, expected to be genetically independent, coexisted in several very rare individuals because neighbouring genes were co-deleted. Cytogenetically visible heterozygous deletions were identified in most cases. We made a series of somatic cell hybrids with the aim of identifying disease-causing genes. Each hybrid carried the deleted homologue of chromosome 11 from different symptomatic donors with cytologically confirmed deletions (van Heyningen et al. 1985). This provided a set of nested deletions allowing us to narrow in on candidate genes predisposing to Wilms' tumour, a childhood kidney malignancy, and aniridia, congenital absence of the iris. We gradually walked along the deletion region, with Wendy Bickmore selecting successive probes to fill the gaps, using different library resources as they became available. A lot of progress was made by collaborating with chromosome-specific groups (Junien and van Heyningen 1990). There was some rivalry, but even more sharing of data and resources. Eventually the gene for Wilms' tumour, the four zinc-finger tumour suppressor *WT1* was cloned in 1990 at Harvard (Call et al. 1990), and for aniridia, the iconic *PAX6* gene was identified in Houston, Texas (Ton et al. 1991). In both cases, we were offered the cDNA clones immediately and for *PAX6* we co-authored the initial paper. We had collected large patient cohorts to study both diseases. We could immediately sequence the *PAX6* cDNA, but defining its genomic organization took time. *PAX6* is not expressed in readily accessible human tissue like blood or skin. Therefore, we identified our earliest human mutations at the mRNA level, using amplification with nested RT-PCR (reverse transcription) from lymphoblastoid cell line (LCL) mRNA (Jordan et al. 1992). We routinely made LCL for all cases collected. There were numerous such "positional cloning" efforts in progress to identify disease-associated genes. Although these projects helped to map disease-relevant regions of the genome bit-by-bit, even after the disease genes had been identified (Couillin et al. 1994), such programmes were not efficient or effective for deriving chromosomal human genome sequence.

Sequencing DNA—revolutionary progress

The arrival of DNA sequencing technology over 40 years ago was critical for triggering the era of the genome. In 1977, two papers appeared a few months apart in the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) (Maxam and Gilbert 1977; Sanger et al. 1977) with different methods. The ingenious Sanger technique, which used radioactively labelled chain terminating inhibitors of nucleotide elongation in polymerase-catalysed primer extension, followed by polyacrylamide gel electrophoresis at single nucleotide sequence resolution, was eventually adopted for most early sequencing applications. The first small genome sequenced was the 5.4 kb bacteriophage ΦX174 (Sanger et al. 1978). From then on, constant innovation and step-by-step improvement of the chemical and engineering technologies allowed

increasingly efficient and accurate sequencing. The history of this progression is described in a review celebrating 40 years of continuing innovation (Shendure et al. 2017). The progress has been breath-taking both technically and in terms of cost reduction. Long-read single-molecule sequencing is now revolutionizing the field (literally, you can sequence in a field), though short read sequencing is still heavily used routinely.

Genome assembly

Genome assembly is still a complex task despite the huge advances in sequencing and in the computational assembly of overlapping fragments (Kolmogorov et al. 2018). Although painfully slow chromosome-specific mapping exercises are no longer needed when genomes in novel species are sequenced, an available framework map can still be very useful and modern technology can refine and confirm it, as shown for the C57BL/6J laboratory mouse genome (Lilue et al. 2018). One tool developed in the mid-1980s for map building is fluorescence in situ hybridisation (FISH). Fluorescently labelled BAC clones were frequently used for hybridisation to denatured metaphase chromosomes and even to interphase nuclei (Lichter et al. 1990). This technology is still used for very large complex genome assemblies, such as the 32 Gb (10× human genome size) axolotl genome (Smith et al. 2018). However, with overlapping sequencing and long-read novel technologies such as Pacific Biosciences (PAC-Bio) and Oxford Nanopore to bridge gaps, genome assembly is becoming easier and more reliable. An important finding for human genomes was the high frequency of copy-number variants, quite a few of which are found to be pathogenic, some associated with specific disease.

Data handling and analysis: bioinformatics

Intertwined with the wet-lab developments, a parallel history of essential discovery and consolidation in data handling and analysis, is also summarized in the history of sequencing (Shendure et al. 2017). The creation of public sequence data repositories, such as ENSEMBL and the University of California Santa Cruz (UCSC) genome browser and their expert curation, were essential steps in establishment of the current genome era. Other genomic data sites include Functional Annotation of the Mammalian Genome (FANTOM) GenomeNet in Japan, Genoscope in France. Sequence alignment algorithms were critical tools for assembling contigs and for evolutionary comparisons. A database (Genome Aggregation Database, gnomAD) for genome-wide variant frequencies in large non-disease populations has been made available to help assess likely pathogenicity of patient variants (Lappalainen et al. 2019). New tools have emerged more recently for gene expression profiles in different tissues (e.g. genotype-tissue expression, GTEx), and indeed in individual cells from whole embryos and organs. The Human Cell Atlas (HCA) project is an exciting current venture, another international collaborative programme, co-led by two women scientists, one in the United States (Aviv Regev) and one in the United Kingdom (Sarah Teichmann) (Regev and Teichmann 2017).

Caenorhabditis elegans—the first multicellular organism genome

I recall with great clarity Sydney Brenner visiting the Genetics Department in Cambridge when I was in the Part II class (1967/8). The department was then still not within the central science sites, but out on the Milton Road. Sydney told us that he was setting out to study a small worm with around 1000 cells because he wanted to understand the development of the nervous system. It seemed amazing and wonderful that he could come up with such a novel plan. John Sulston began this project in 1969 when he joined the Brenner lab. His first task was to study and document the development of the *C. elegans* nervous system under the microscope, using a highly observational approach. Following early success in this endeavor, Sulston and other young scientists arriving at the lab worked out the full cell by cell developmental pathway of the little nematode. The process is engagingly described in John Sulston's Nobel Lecture (Sulston 2003). In parallel, they created many phenotypically defined mutations, including a class termed un-coordinated (*unc*)—movement mutants—a proportion of which were likely to affect the neural system. They were mapped, using conventional crosses, to the six previously defined linkage groups (Brenner 1974). Identifying the mutant genes and deciphering their function was, however, a slow and cumbersome process. It became clear that identification of all *C. elegans* genes needed to be undertaken to explore the biology properly.

As discussed above, the congruence of several new technologies allowed this project to be undertaken—identifying all *C. elegans* genes, and their non-coding environment, was to become the first genome project for a multicellular organism. Alan Coulson who had worked with Fred Sanger until his retirement, was able to join the *C. elegans* genome project. Genome size was initially estimated using DNA reassociation kinetics (Sulston and Brenner 1974). Dense coverage genomic libraries were created to produce a physical map and for the sequencing, initially in lambda phage vectors, but then they switched to cosmids with larger inserts (~ 40 kb). Clones were "fingerprinted" using restriction enzyme digestion followed by electrophoresis to reveal patterns of different sized bands. A

genetic map was collated. The overlapping fragments were assembled into contigs using automated comparison of clones—John Sulston had to teach himself Fortran to produce a programme for this. However, there were far too many contigs because of the many unfilled gaps. Robert Waterston was visiting the Sulston lab at this point, aware that YAC cloning was being developed in Maynard Olson's group in Seattle and decided to try this large insert (200–500+ kb) cloning system to try to bridge the gaps (Coulson et al. 1991). He made a YAC library from the *C. elegans* genome, and succeeded in closing most gaps. The Sulston and Waterston labs, in Cambridge and St Louis, thus embarked on a 20 year harmonious collaboration starting with *C. elegans* genome sequencing. The genome was deemed sufficiently complete for publication in 1998 (The *C. elegans* Consortium 1998). The final reported size, following sequencing, is 100 Mb. About 20,000 genes were defined by the nematode genome, and confirmed by further analysis (Hillier et al. 2005). This is a considerably larger number than predicted and significantly higher than the gene numbers subsequently recognized in Drosophila. The completion of the genome was deemed to be the beginning of a rich era of biological exploration. Genes implicated in mutant phenotypes could now be readily identified and functional studies launched using reverse genetics. Most of the genes identified fulfilled evolutionarily conserved functions. *C. elegans* proved to be an excellent model organism, in terms of biology as well as pointing the way to the genomic era.

In the course of this extended project, many mutually agreed simple rules were devised to allow completely open communications between the partners in this UK–US collaboration, with virtually immediate public release of sequence data, while also ensuring that the scientists involved received due recognition for their labours (Sulston and Ferry 2002; Sulston 2003; Maxson Jones et al. 2018).

The human genome project

Mapping of human genes had been in progress for some time before sequencing technology was discovered. Some of the history of this early era is described in the proceedings of a discussion meeting held in 2014 (Jones 2015) under the aegis of The History of Modern Biomedicine (http://www.histmodbiomed.org/witsem/vol54.html). There were extensive chromosome-specific gene and disease/phenotype maps developed in many labs internationally, and consolidated at regularly convened Human Gene Mapping workshops between 1973 and 1991 (e.g. Junien and van Heyningen 1990). After 1991, as the volume of mapping data increased, mapping meetings became chromosome-specific gatherings under the auspices of the Human Genome Organisation, HUGO, founded in 1989 (http://www.hugo-international.org/history). The methodology continued to evolve during this era. A significant number of major genes mutated in Mendelian diseases were identified in the 1980s and 1990s (Collins 2003), before the systematic sequencing of the whole human genome had begun in earnest.

Well before the worm genome was finished, there was serious discussion about the possibility of sequencing the human genome. Discussions ranged from political through tactical to scientific and there were complex negotiations about dividing up the work and funding. The story has been told from several viewpoints including in some detail by John Sulston (Sulston and Ferry 2002), whose leading role in HGP has been widely acknowledged. Sulston's narrative is easy to read, but is definitely from his own view-point. There are of course other somewhat different perspectives on the historical sequence of events and on the agreements and disagreements (Collins et al. 2003). The key decision for US participation was taken by influential scientists and science funding agencies, at a Banbury Center meeting in 1989 (see Green et al. 2015). The exact approaches to delivering the human genome project evolved continuously on both sides of the Atlantic. In addition, there were important but smaller contributions from several other countries, including France, Germany, China and Japan. The ultimate aim was to deliver a genetic and a physical map prior to embarking on the full sequence. Having strong commitment to funding the project both in the United States and the United Kingdom helped to keep the resolve going for such a huge and long-term effort. In the United Kingdom, the Wellcome Trust and the Medical Research Council (MRC) were involved; in the United States, the two main funders were the Department of Energy (DOE) and the National Institutes of Health (NIH). However, commercial organisations, particularly in the United States, were also lining up to participate, as the potentially saleable usefulness of multiple aspects of HGP became clear. There were several attempts to sequester some or all of the data behind privatized pay walls. Craig Venter, initially an NIH intramural scientist, was unhappy about the level of resources available to him within NIH, so he resigned and co-founded a non-profit company called TIGR (The Institute for Genomic Research), where in 1991 he claimed to be the initiator of a project to sequence cDNA clones (Expressed Sequence Tags, ESTs), which he proposed to patent for biomedical discovery use (Sulston and Ferry 2002). As noted earlier, others had also suggested EST mapping as a first step in HGP. cDNA mapping is faster but it is not possible to gather all cDNAs corresponding to all the genes, and people were also becoming aware that regulation of gene expression generally lies within non-coding regions of the genome. Nevertheless, Human Genome

Sciences, a subsidiary company of TIGR, did license ESTs for privileged early commercial access. Rescue arrived in 1994, when Merck, the pharmaceutical company, agreed to support the EST project, with no strings attached, encouraging immediate public release as sequence was generated.

By 1996, John Sulston and Robert Waterston, the leaders of the "worm project" in the United Kingdom and the United States, respectively, felt confident that working with a broader and more international consortium, would allow the full HGP, 30-fold larger than *C. elegans*, to be delivered by 2005. The main participants met that February in Bermuda, to thrash out the principles that were to be followed: despite one or two detractors who feared quality problems, there was strong agreement that immediate daily release of raw sequence data, available for full public access, must be implemented as a condition of consortium membership. More subtle details of the history around this era have been published (Maxson Jones et al. 2018). The successful *C. elegans* scheme, of building a minimum tiling path of BAC clones for sequencing, was accepted as a good functional starting strategy. YAC libraries could sometimes be used to bridge gaps. Much of the effort built on the early chromosome-by-chromosome disease-mapping work.

In May 1998, just as everyone was arriving at the annual Cold Spring Harbor Human Genome meeting, Craig Venter announced that his new enterprise, Celera Genomics, could complete the human genome project without other collaborators using solely shotgun sequencing and overlap assembly algorithms. Financial and technological support would come from partners Applied Biosystems (ABI), the company that was fast developing novel sequencing machines. He would prove Celera's capacity to achieve this by quickly delivering the 138 Mb Drosophila genome—following quick negotiations with fly geneticist Gerry Rubin, who was leading the fly project.

Initially, this announcement of the Celera plans produced doubt—about the ability of Venter and his technique to deliver high quality finished sequence; dismay—about the re-surfacing of plans for privatization of at least some of the data; and consternation—about how the funding agencies, mainly in the United Kingdom and the United States, would view continuing and even enhanced support for the public project (Sulston and Ferry 2002). The public HGP continued to be supported and the two projects moved forward extraordinarily fast, so that initial draft sequences of the human genome were published simultaneously in rival journals in 2001: the Celera results were published in Science (Venter and Celera Genomics 2001), while the collaborative HGP sequence emerged in Nature (The International Human Sequencing Consortium 2001). The completion of the human genome was announced in April 2003 amid great fanfare, with representatives of both the public and the Celera projects present (https://www.genome.gov/11006929/2003-release-international-consortium-completes-hgp/) (https://www.nature.com/news/2003/030414/full/news030414-1.html). The timing coincided with the 50th anniversary of the discovery of the structure of DNA. From that time to 2006, we saw the publication by the public project of the sequence and organisation of each of the 22 human autosomes and the X and Y chromosomes, starting with the smallest and ending with the largest, chromosome 1 (Gregory et al. 2006). The controversies between the shotgun sequencers and the mapper-sequencers have continued to be discussed and debated (Waterston et al. 2003; Istrail et al. 2004).

The post-HGP landscape and lessons learned

Since the early results of HGP a vast genome industry has evolved globally. One of the early steps after the "reference genome" was defined was to explore variability among "normal" individuals in the 1000 genomes project (Abyzov et al. 2015). Human genome structure and function continue to be extensively elaborated and the United Kingdom continues to make important contributions, with amazing longitudinal cohort studies (ALSPAC and the UK Biobank), which gradually turned into genome projects as large-scale genome sequencing became feasible. UK10K (Rare Genetic Variants in Health and Disease) and DDD (Deciphering Developmental Disorders) were vanguard genome projects for cohorts with undiagnosed congenital abnormalities. The still bolder 100,000 Genomes project has now completed the sequencing phase and the aim has been expanded to 5 million genomes.

Multiple aspects of genome regulation have been explored, producing iterations of the Encyclopaedia of DNA Elements (ENCODE) (The Encode Project Consortium et al. 2007; Diehl and Boyle 2016). Studying chromatin organization has been a major preoccupation of many groups worldwide, for example, exploring relatively short-range DNA interactions using ingenious chromatin conformation capture technologies (Dixon et al. 2012; Nora et al. 2013). My own group had long been interested in regulatory control of developmental gene expression (Kleinjan and van Heyningen 2005) and in disease-associated regulatory changes (Fantes et al. 1995; Bhatia et al. 2013, 2015). Our understanding of the effects of genomic variation among humans has grown exponentially over the past few years, with the much quicker and cheaper sequencing technologies available. Ever-growing

population sequencing studies are carried out on healthy (Elliott et al. 2018) and disease cohorts (Fitzgerald et al. 2014; Turnbull et al. 2018), shedding light, increasingly, on the mechanisms.

Since the beginning of the genome era, many vertebrate and invertebrate genomes have been sequenced and deposited in public databases such as Ensembl (http://www.ensembl.org/info/about/species.html) by countless groups world-wide since the first multicellular organism, *C. elegans*. Comparative analysis of these has brought fantastic insights into evolution and functional mechanisms. Not to be left behind, microorganisms and virus genomes have also generated huge excitement over the past decade (e.g. http://bacteria.ensembl.org/species.html). The significance of microbiome diversity is widely considered not only in learned journals, but even in newspapers and magazines. These days, genome analysis is the first response to infectious disease outbreaks globally. New sequencing technologies, which can be carried out in the field on locally collected samples, are revolutionizing this approach.

Science and society

It is truly an exciting and exhilarating privilege to be living in this fast-paced era of genomics. I have barely been able to convey the history of one small but critical corner. It is unbelievable how much thinking and ingenious technology has been developed and in some cases already discarded as new approaches have taken over. The history of the development of sequencing, both proteins and nucleic acids, illustrates this most tellingly. Two Nobel Prize-winning technologies (1958, proteins; 1980, DNA—both awarded to Fred Sanger but the latter shared with Paul Berg and Walter Gilbert) have been transformed almost out of recognition and are no longer used routinely in their original format, but mass spectroscopy for sophisticated protein analysis and "next generation sequencing" as well as single molecule long-range DNA sequencing are now used very widely and increasingly cheaply. The virtual disappearance of Southern blotting is another example of the fast turnover of hot technology.

Throughout the history of the genome era there have been controversies and major ethical issues to consider. In some cases we have developed ways of dealing with the problems, at least temporarily and in ways that seem acceptable today, but today's solutions may not remain satisfactory in future. Perhaps for the first time in history there has been a parallel thread of an ELSI (ethical, legal, societal implications) programme along with the genome programme. There was much debate about consent issues and different aspects of human rights, for example, about the ethics of what results should and should not be revealed to people at different stages of their life. Thought also has been given to the ownership of tissue samples. I am not sure we have fully solved this and there are real problems, in my view, about trying to apply today's new rules to yesterday's cases. There remain conflicts, in many cases, between individual rights and societal rights. But the main thing is that we are thinking about these matters and discussing them, almost obsessively. One of the great champions of public ownership of data and information was, undoubtedly, John Sulston. I knew him fairly well in the first decade of this century, when we both sat on the Human Genetics Commission, under the chairmanship of human rights lawyer Helena Kennedy. It was a heady exciting time. Also with us in that group we had Medical Law professor and emerging author Alexander McCall Smith, and the sickle cell disease expert Professor of Nursing, Elizabeth Anionwu. At times, there was overkill on rules: for a while it was suggested that cancer tissues donated for research with consent, could not be used after the donor died—obviously a nonsense ruling. But we do need to think constantly about what is fair and reasonable at all times and to re-think these ideas as technology and society's outlook evolves. The Sulston voice was always well considered and logical and sought to maintain the balance between individual rights and societal needs. However, there are always sins of omission as well as commission, and in fact the boundary between these is very slight. We do now realize that the genomes studied are not proportionately representative of the world's population. We need to enable the study of genomes within Africa, parts of Asia and South America and also among indigenous populations scattered around the world. Not only will we learn much important biology from this, but we shall also be in a better position to deliver much more equitably the benefits of genomic studies to everyone, as the rich diversity is understood and utilised. However, it must be borne in mind that not everyone will consent to have their genome studied. These points need to be discussed continuously, as views do change. The information revolution and big-data developments have, of necessity, accompanied the genomic flood. The influence of these advances will be yet wider than even the genome revolution itself. We need to remember that the saying: "May you live in interesting times" is considered a curse in some ways, but it is our task to ensure that the horrors are avoided and as much as possible of the benefits realized.

Compliance with ethical standards

Conflict of interest This essay is dedicated to the memory of both John Sulston, who died just before the Genetics Society Centenary project was launched, and Sidney Brenner, who died the day after submission. Both were thoughtful, clever, inventive and principled scientists, and role models. The author declares that she has no conflicts of interest.

References

Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF et al. (2015) Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat Commun 6:7256

Anand R, Villasante A, Tyler-Smith C (1989) Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed-field gel electrophoresis. Nucleic Acids Res 17:3425–3433

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA 74:3171–3175

Bhatia S, Bengani H, Fish M, Brown A, Divizia MT, de Marco R et al. (2013) Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. Am J Hum Genet 93:1–9

Bhatia S, Gordon CT, Foster RG, Melin L, Abadie V, Baujat G et al. (2015) Functional assessment of disease-associated regulatory variants in vivo using a versatile dual colour transgenesis strategy in zebrafish. PLoS Genet 11:e1005193

Bigger C, Murray K, Murray NE (1973) Recognition sequence of a restriction enzyme. Nat New Biol 244:7–10

Bobrow M, Cross J (1974) Differential staining of human and mouse chromosomes in interspecific cell hybrids. Nature 251:77–79

Borck K, Beggs J, Brammar W, Hopkins A, Murray N (1976) The construction in vitro of transducing derivatives of phage lambda. Mol Gen Genet 146:199–207

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Brenner S (1974) The genetics of *Caenorhabditis elegans*. Genetics 77:71–94

Call KM, Glaser T, Ito CY, Buckler AJ, Pelletier J, Haber DA et al. (1990) Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell 60:509–520

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell 12:1–8

Collins FS (2003) Positional cloning moves from perditional to traditional Francis. Nat Genet 9:347–350

Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: Lessons from large-scale biology. Science 300:286–290

Couillin P, Le Guern E, Vignal A, Fizames C, Ravise N, Delportes D et al. (1994) Assignment of 112 microsatellite markers to 23 chromosome 11 subregions delineated by somatic hybrids: comparison with the genetic map. Special Issue: Assignment of 112 microsatellite markers to 23 chromosome 11 subregions delineated by somatic hybrids. Genomics 21:379–387

Coulson A, Kozono Y, Lutterbach B, Shownkeen R, Sulston J, Waterston R (1991) YACs and the *C. elegans* genome. BioEssays 13:413–417

Diehl AG, Boyle AP (2016) Deciphering ENCODE. Trends Genet 32:238–249

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376–380

Elliott L, Sharp K, Alfaro-Almagro F, Shi S, Miller K, Douaud G et al. (2018) Genome-wide association studies of brainimaging phenotypes in UK Biobank. Nature 562:210–216

Fantes J, Redeker B, Breen M, Boyle S, Brown J, Fletcher J et al. (1995) Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. Hum Mol Genet 4:415–422

Fitzgerald TW, Gerety SS, Jones WD, van Kogelenberg M, King Da, McRae J et al. (2014) Large-scale discovery of novel genetic causes of developmental disorders. Nature 519:223–228

Green ED, Watson JD, Collins FS (2015) Twenty-five years of big biology. Nature 526:29–31

Gregory SG, Barlow KF, McLay KE, Kaul R, Swarbreck D, Dunham A et al. (2006) The DNA sequence and biological annotation of human chromosome 1. Nature 441:315–321

Hillier LW, Coulson A, Murray JI, Bao ZR, Sulston JE, Waterston RH (2005) Genomics in *C elegans*: So many genes, such a little worm. Genome Res 15:1651–1660

Holland J, Coffey AJ, Giannelli F, Bentley DR (1993) Vertical integration of cosmid and YAC resources for interval mapping on the X-chromosome. Genomics 15:297–304

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R et al. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci 101:1916–1921

Jeffreys AJ, Craig IW, Francke U (1979) Localisation of the G gamma-, A gamma-, delta- and beta-globin genes on the short arm of human chromosome 11. Nature 281:606–608

Jeffreys AJ, Monckton DG, Tamaki K, Neil DL, Armour JAL, MacLeod A et al. (1993) Minisatellite variant repeat mapping: Application to DNA typing and mutation analysis. Exp Suppl 67:125–139

Jones EM (2015). Human gene mapping workshops c.1973–c.1991. In: Jones EM, Tansey EM (eds) Wellcome witnesses to contemporary medicine, Vol. 54. Queen Mary University of London: London

Jordan T, Hanson I, Zaletayev D, Hodgson S, Prosser J, Seawright A et al. (1992) The human PAX6 gene is mutated in two patients with aniridia. Nat Genet 1:328–332

Junien C, van Heyningen V (1990) Report of the committee on the genetic constitution of chromosome 11. Cytogenet Cell Genet 55:153–169

Kao F, Jones C, Puck TT (1977) Genetics of cell-surface antigens: regional mapping of three components of the human cell-surface antigen complex, AL, on chromosome 11. Somat Cell Genet 3:421–429

Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: Emerging mechanisms and disruption in disease. Am J Hum Genet 76:8–32

Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F et al. (2018) Chromosome assembly of large and complex genomes using multiple references. Genome Res 28:1720–1732

Lappalainen T, Scott AJ, Brandt M, Hall IM (2019) Genomic analysis in the age of human genome sequencing. Cell 177:70–84

Lichter P, Tang C-JC, Call K, Hermanson G, Evans GA, Housman D et al. (1990) High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. Science 247:64–69

Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R et al. (2018) Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet 50:1574–1583

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Maxson Jones K, Ankeny RA, Cook-Deegan R (2018) The Bermuda Triangle: The pragmatics, policies, and principles for data sharing in the history of the human genome project. J Hist Biol 51:693–805

Morton NE (1996) Logarithm of odds (lods) for linkage in complex inheritance. Proc Natl Acad Sci USA 93:3471–3476

Nora EP, Dekker J, Heard E (2013) Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? BioEssays 35:818–828

Pontecorvo G (1971) Induction of directional chromosome elimination in somatic cell hybrids. Nature 230:367–369

Regev A, Teichmann S (2017) The Human Cell Atlas: from vision to reality. Nature 550:451–453

Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL et al. (1978) The nucleotide sequence of bacteriophage φX174. J Mol Biol 125:225–246

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA et al. (2017) DNA sequencing at 40: Past, present and future. Nature 550:345–353

Smith JJ, Timoshevskaya N, Timoshevskiy VA, Keinath MC, Hardy D, Voss SR (2018) A chromosome-scale assembly of the Aaxolotl genome. Genome Res 29:317–324

Solomon E, Bodmer WF (1979) Evolution of sickle cell variant gene. Lancet 8122:923

Southern EM (1975) Detection of specific sequences among DNA fragments. J Mol Biol 98:503–517

Sulston JE (2003) Caenorhabditis elegans: The cell lineage and beyond (Nobel lecture). ChemBioChem 4:688–696

Sulston JE, Brenner S (1974) The DNA of *Caenorhabditis elegans*. Genetics 77:95–104

Sulston J, Ferry G (2002) The common thread. Bantam Press, London

The C. elegans Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282:2012–2018

The Encode Project Consortium, Holmes I, Löytynoja A, Karolchik D, Frankish A, Foissac S et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

The International Human Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Ton CCT, Hirvonen H, Miwa H, Weil MM, Monaghan P, Jordan T et al. (1991) Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region. Cell 67:1059–1074

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Boardman Pretty F et al. (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ 361:1687

van Heyningen V, Boyd PA, Seawright A, Fletcher JM, Fantes JA, Buckton KE et al. (1985) Molecular analysis of chromosome 11 deletions in aniridia-Wilms tumor syndrome. Proc Natl Acad Sci USA 82:8592–8596

van Heyningen V, Craig IW, Bodmer WF (1973) Genetic control of mitochondrial enzymes in human-mouse somatic cell hybrids. Nature 242:509–512

Venter JC, Celera Genomics (2001) The sequence of the human genome. Science 291:1304–1351

Wang DG, Fan J, Siao C, Berno A, Young P, Sapolsky R et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wang M, Chen X, Shouse S, Manson J, Wu Q, Li R et al. (1994) Construction and characterization of a human chromosome 2-specific BAC library. Genomics 24:527–534

Waterston RH, Lander ES, Sulston JE (2003) More on the sequencing of the human genome. Proc Natl Acad Sci USA 100:3022–3024.